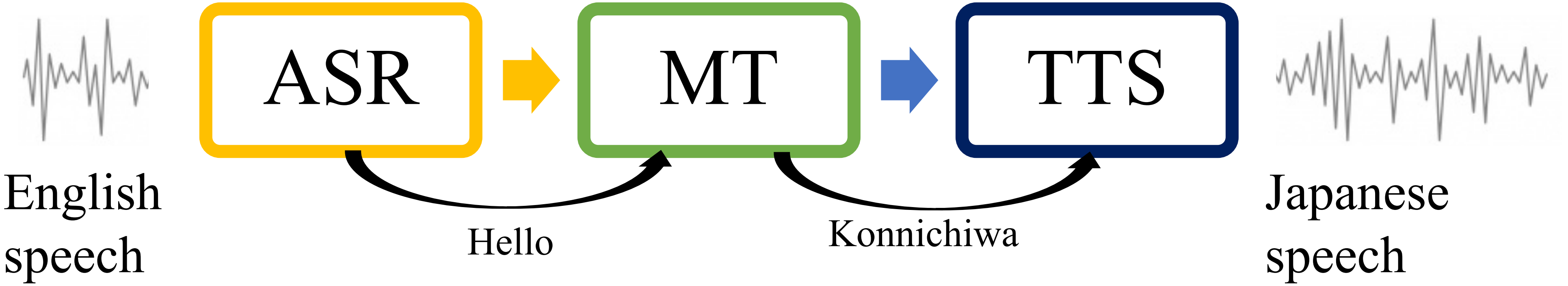
Transformer-based Direct Speech-to-speech Translation with Transcoder

Takatomo Kano[1], Sakriani Sakti[1,2], and Satoshi Nakamura[1,2]

1. Nara Institute of Science and Technology, Japan

2. RIKEN, Center for Advanced Intelligence Project AIP, Japan

Introduction



The traditional approach solves the speech-to-speech translation step by step.

However, there are many limitations.

- ✓ Cascade of ASR, MT, and TTS using text.
- ✓ All of which is independently trained and tuned.

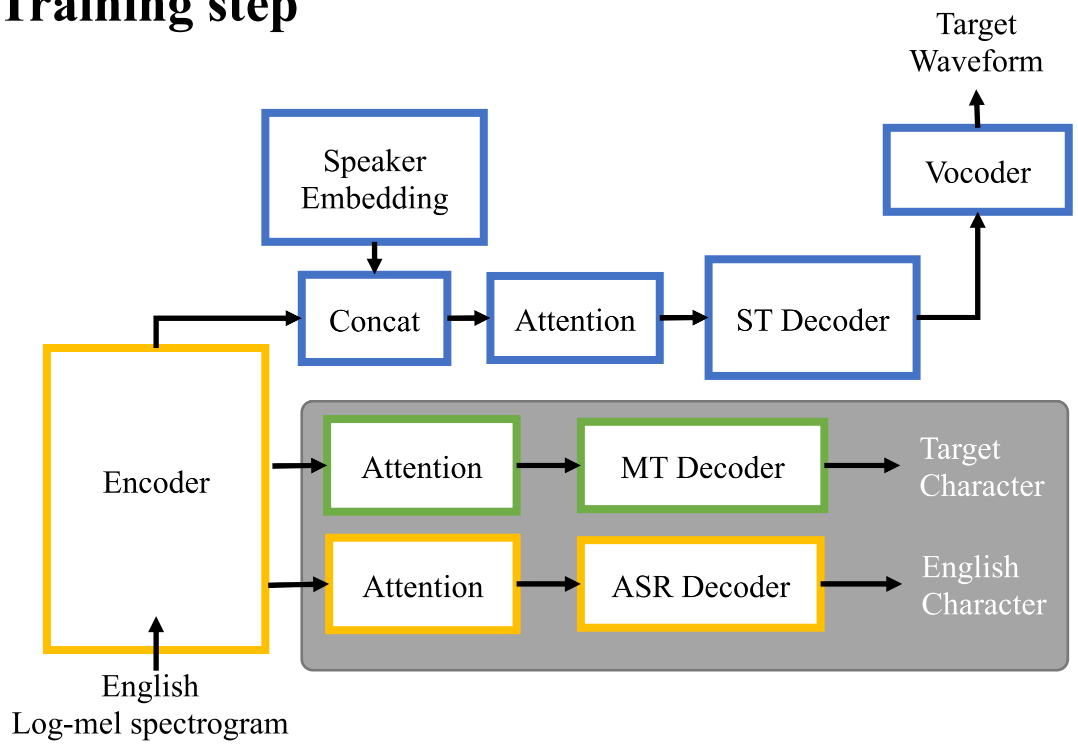
Existing work: Multi-task Speech-to-Speech Translation[Ye et al., 2019]

Multi-task model

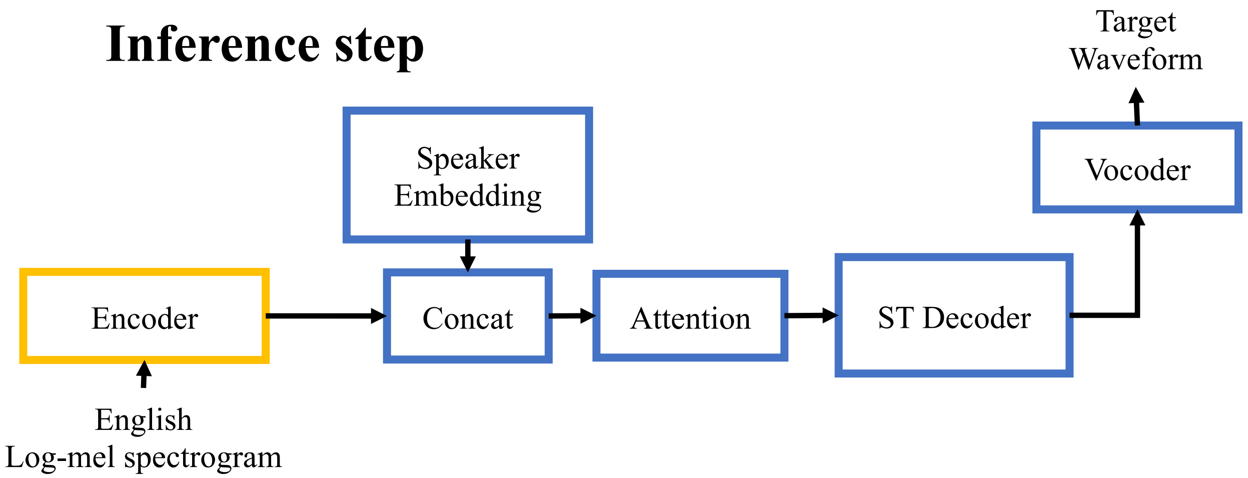
✓ **The model has three decoders but uses only one decoder for translation.**

Multi-task model
[Weiss et al., 2017, Ye et al., 2019]

Training step



Inference step

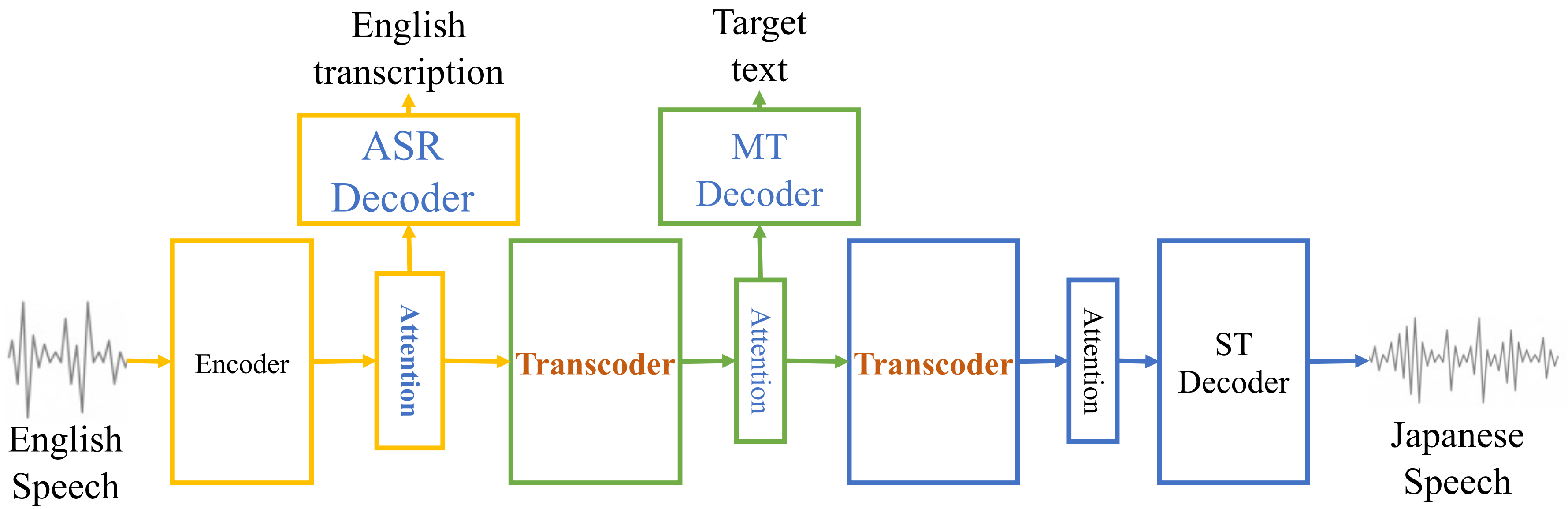


- ✓ **The model uses one attention module for alignment source and target speech.**
- ✓ **ASR and MT decoders do not work in the inference step.**
  - ✓ **There is no information sharing from ASR and MT.**
- ✓ **All modules use RNN.**

Proposed: Transformer-based Direct Speech-to-speech Translation with Transcoder

Proposed

✓ **The model utilizes all trained modules to translate.**



- ✓ **ASR and MT decoders provide attention alignment necessary in the next block.**
  - ✓ The model uses a combination of 3 attentions for alignment source to target speech.
- ✓ **Transcoder converts different hidden states** in speech translation.
  - ✓ The model converts ASR and MT hidden states into MT and TTS hidden states.
- ✓ We revamp the **RNN**-based speech-to-speech translation model by the **Transformer**.

Experiment Results

BLEU and METEOR scores of speech-to-speech translation

Model	Syntactic similar				Syntactic distant			
	En to ES		Ja to Ko		En to Ja		Ja to En	
	BLEU		METEOR		BLEU		METEOR	
Baseline: Cascade (RNN)	38.9	47.7	38.7	49.1	32.5	44.2	32.0	43.2
Baseline: Cascade (Transformer)	41.3	52.1	41.0	51.1	34.1	45.2	35.0	45.3
Multi-task (RNN)	38.8	48.2	39.1	49.9	33.2	45.5	34.2	45.0
Multi-task (Transformer)	43.1	58.8	42.5	58.3	36.9	52.6	38.3	48.4
Transcoder(Transformer)	44.0	59.3	42.9	58.8	40.6	56.6	41.0	55.8

✓Our proposed approach outperforms Multi-task speech translation.

✓The proposed method with Transformer further improves performances.

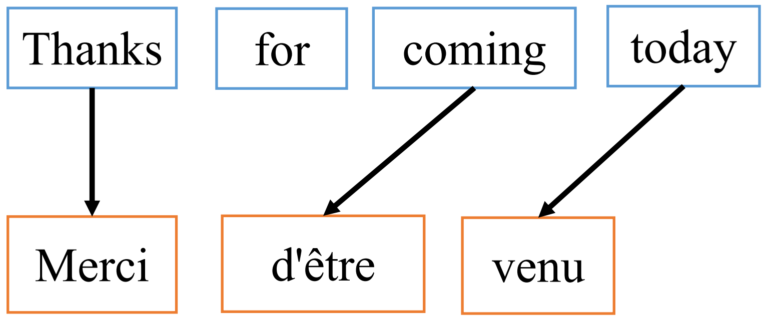
End of my highlight talk.

Limitations in Existing Approach

Many existing works only perform on syntactically similar language pairs.

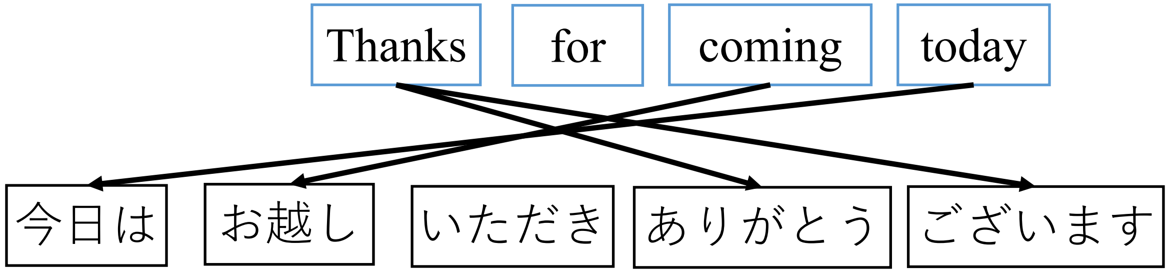
- Syntactically similar language pairs have
  - Similar word ordering(SVO-SVO or SOV-SOV)
  - A lot of one-to-one monotonic alignments.
    - English-French, English-Spanish, Japanese-Korean

English to French translation



- Syntactically distant language pairs have
  - Different word ordering(SVO-SOV)
  - A lot of many-to-many alignments.
    - English-Japanese
  - Difficult to translate

English to Japanese translation



Direct End-to-End Speech Translation for Distant Language Pairs

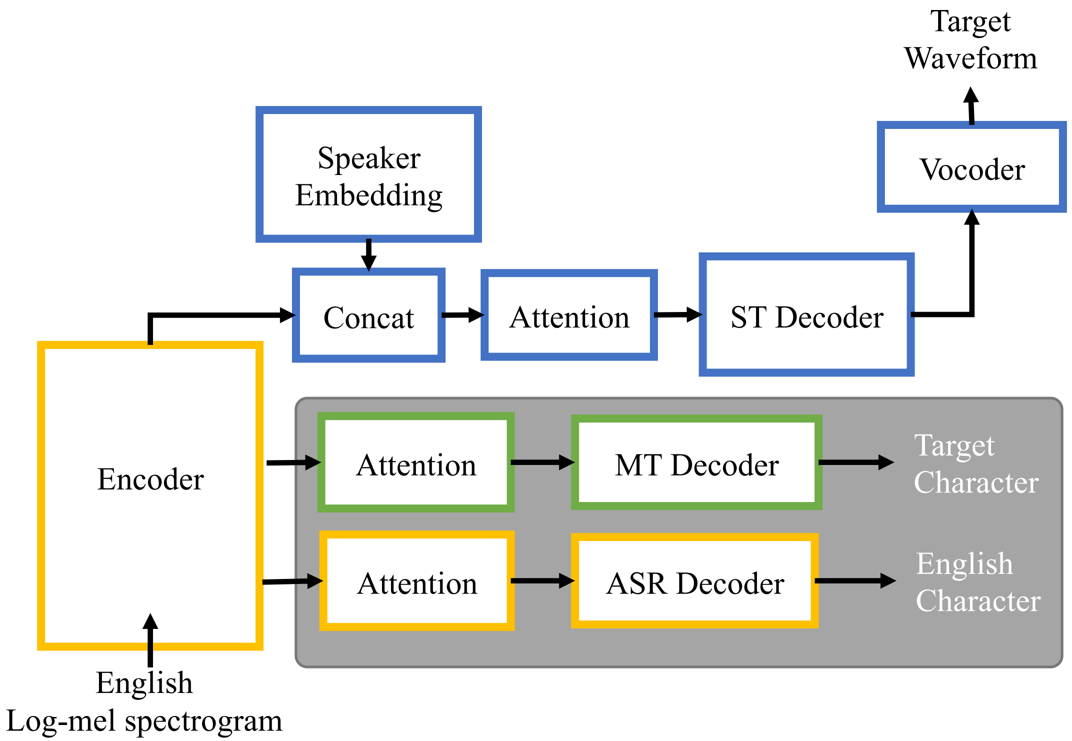
Existing work: Multi-task Speech-to-Speech Translation[Ye et al., 2019]

The speech encoder states are most important.

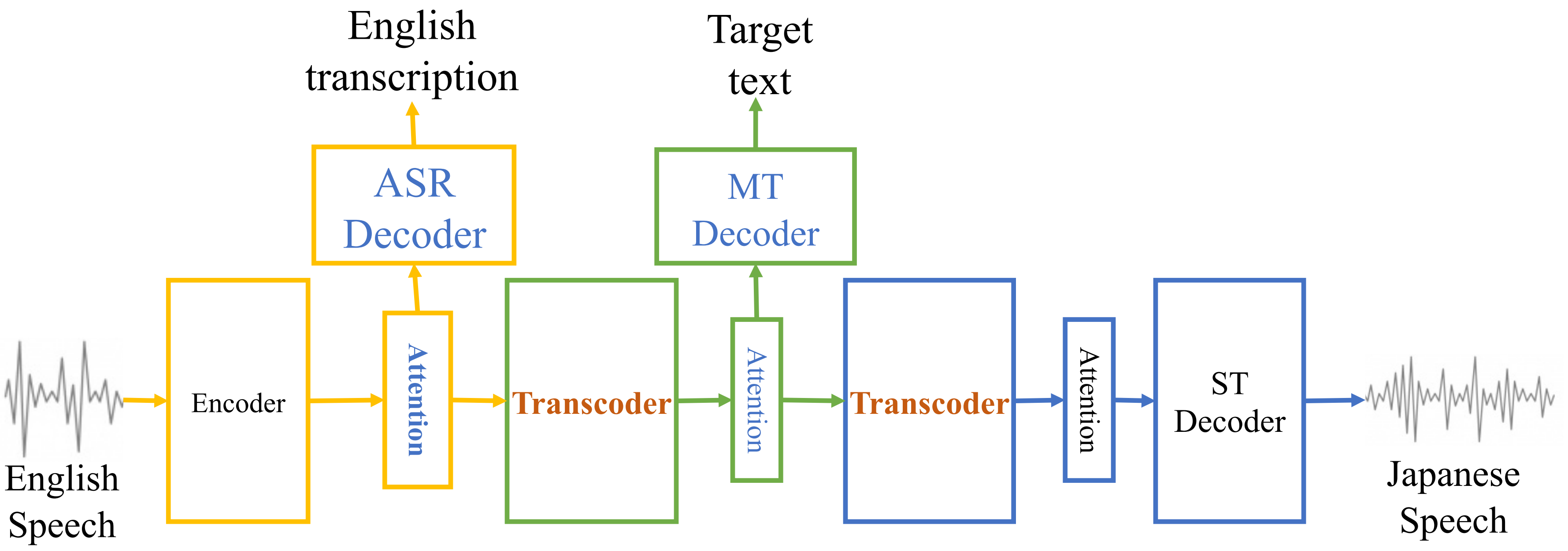
- ✓ Source text decoder helps to tune source speech encoder states.
- ✓ Source text helps find the right segment on source speech states.

Limitations

- ✓ Use a simple encoder-decoder model to translate input speech to target speech.
- ✓ It's challenging to handle difficult translation tasks due to simple architecture.
- ✓ All modules use RNN.



Proposed: Transformer-based Direct Speech-to-speech Translation with Transcoder



- ✓ **ASR and MT decoders provide attention alignment necessary in the next block in the inference step.**
  - ✓ The model uses a combination of 3 attentions for alignment source speech to target speech.
- ✓ **Transcoder converts different hidden states** in speech translation.
  - ✓ The model converts ASR and MT hidden states and MT and TTS hidden states.
- ✓ We revamp the **RNN**-based speech-to-speech translation model with the **Transformer**.

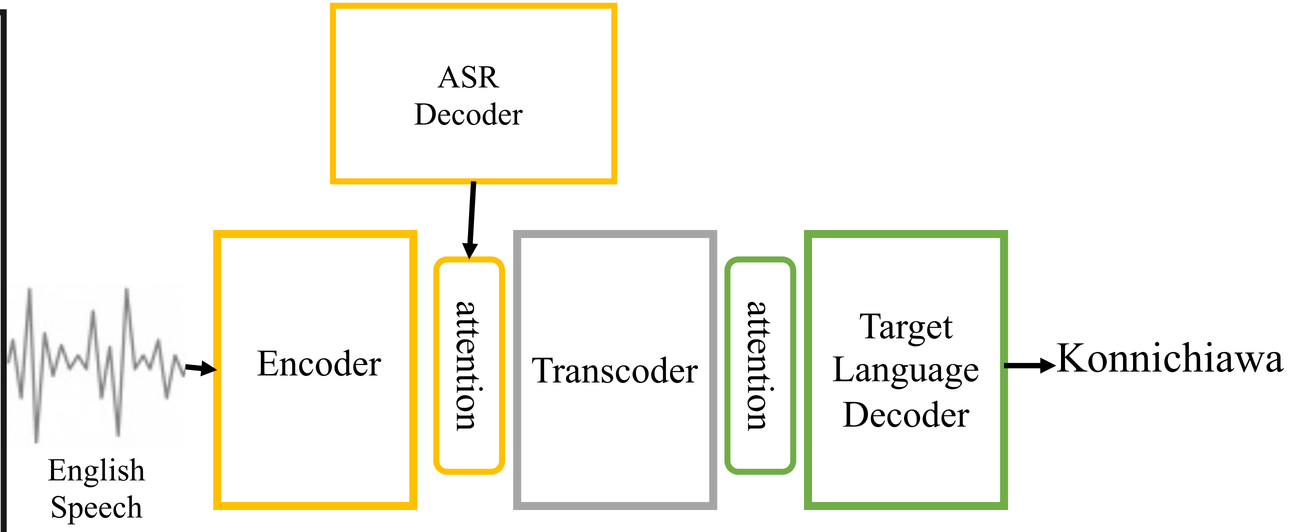
Proposed: Transformer-based Direct Speech-to-speech Translation with Transcoder

Alignment information is most important for Speech Translation.

- ✓ This model provide attention alignment necessary in the next block .
- ✓ Transcoder converts different hidden states in speech translation.

Limitations

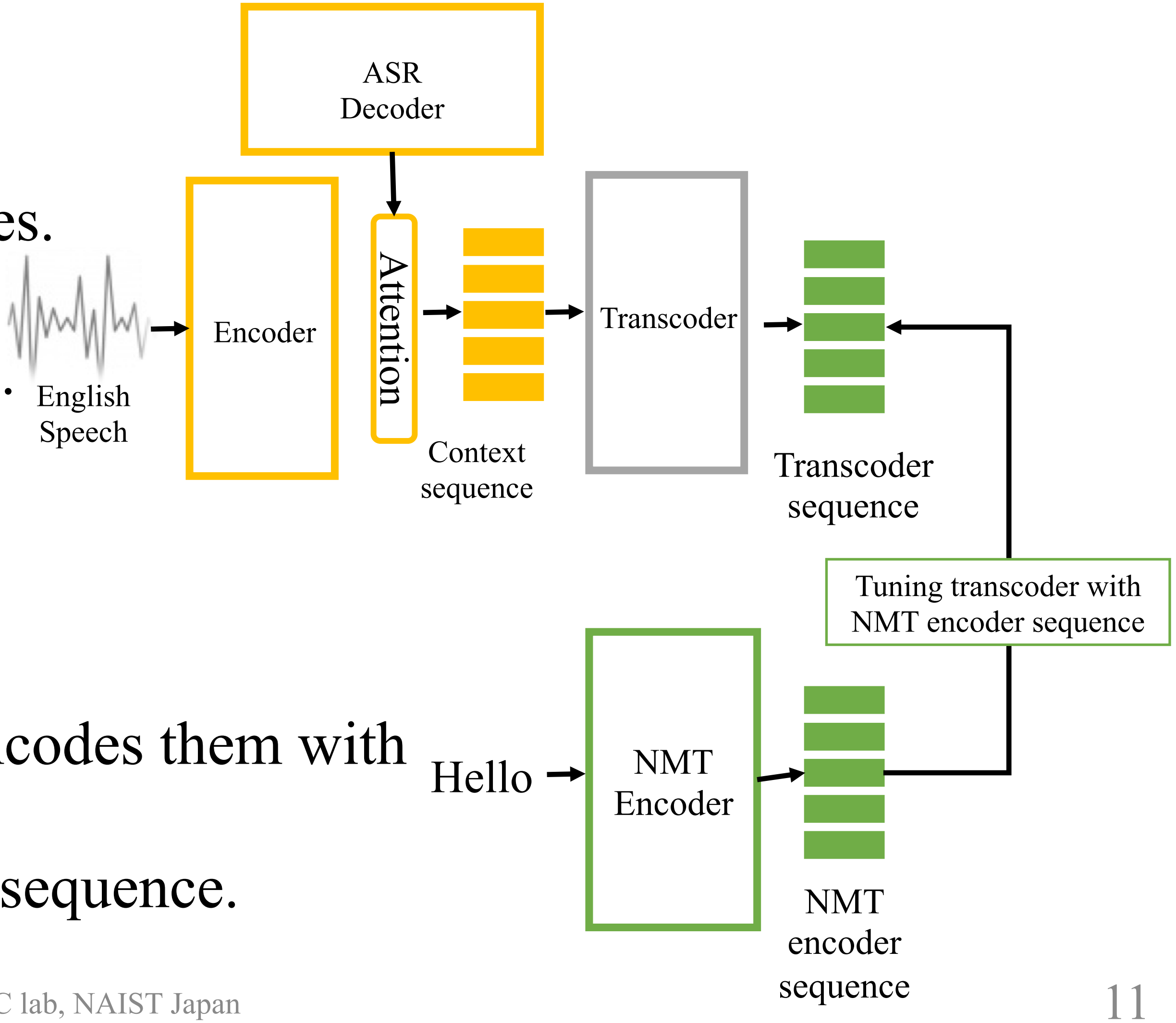
- ✓ The model has deep architecture, and the training is difficult.
- ✓ We need additional training to convert ASR hidden states into MT hidden states.



Proposed:
Training mechanism for Transcoder

The difference between ASR and MT.

- ✓ ASR models input acoustic phoneme sequences.
- ✓ MT models input word sequence.
- ✓ Each hidden states have different distributions.



Transcoder's training

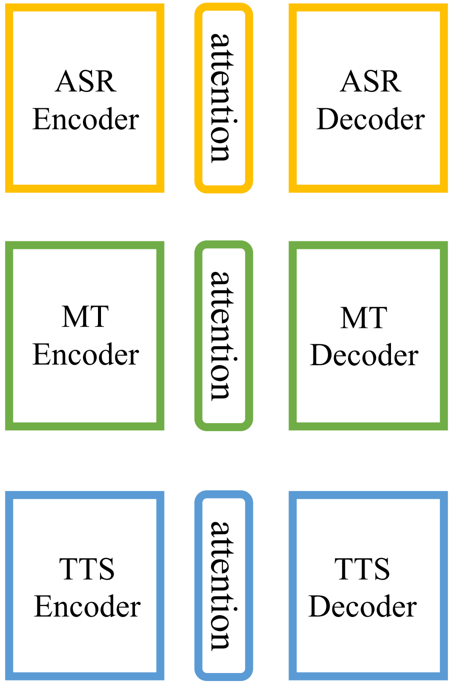
- ✓ Transcoder gets ASR context sequence and encodes them with considering context information.
- ✓ A pre-trained MT encoder provides the target sequence.

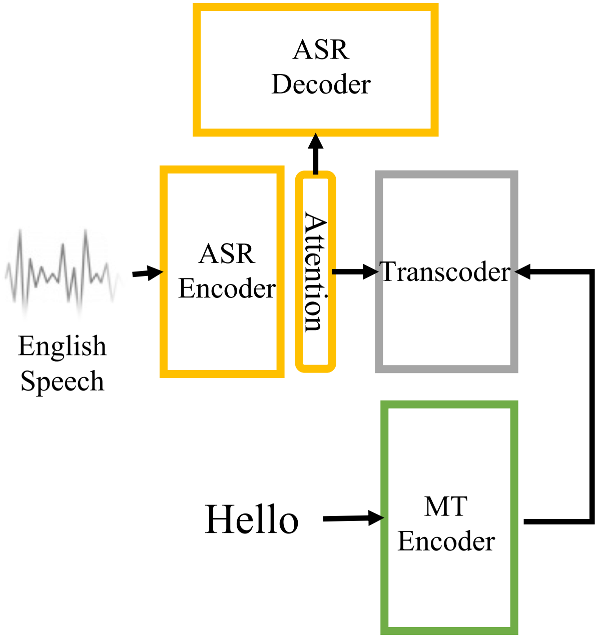
Proposed: Training mechanism for Transcoder

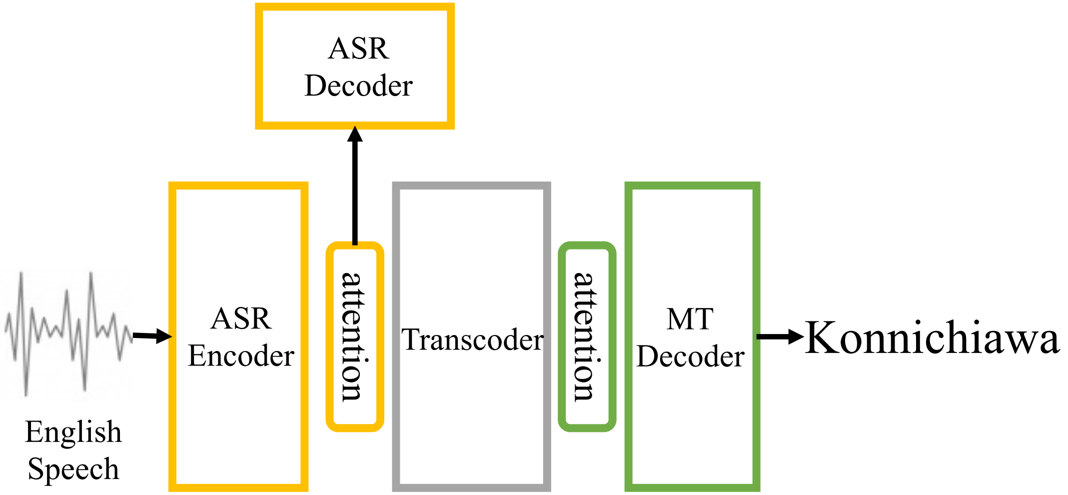
We extend the architecture and train model step by step.

- ✓ We get an idea from Curriculum Learning. We change the model architecture and task difficulty.
- ✓ During the training, the model becomes deep, and the task becomes difficult at each step.
- ✓ We utilize a pre-trained MT and TTS encoder as a teacher model.

Step1: Pre-train | Step2: Train Transcoder | Step3: Extend MT decoder | Step4...







Experimental setups

Transformer setting	
Encoder layers	3
Decoder layers	6
Multi-head	8
Transformer hidden size	256
Transformer FFN [7] hidden size	1024
RNN setting	
Encoder layers	4
Decoder layers	2
RNN type	LSTM
Multi-head	8
RNN hidden size	256
Speech input and output layer setting	
Speech input size	80 (mel) * 3 (frames)
Speech mel out size	80 (mel) * 5 (frames)

Optimizer setting	
Warm-up steps for Transformer	8000
Optimizer method	Adam
Learning rate decay	0.8
Training steps	2,000,000
Batch size (ASR / MT / TTS / ST)	64 / 128 / 32 / 16

- ✓ Input is Japanese or English natural speech.
- ✓ Target is Japanese, English, Korean, or Spanish generated speech.

- ✓ We use the ASR to transcribe the generated Mel sequences and evaluate translation performance with BLEU and METEOR.

Experimental results

We evaluate ASR and MT performance to compare.

- ✓ RNN and Transformer model.
- ✓ Syntactically similar and distant language.

ASR WER

Model	Natural English	Natural Japanese	Generated English	Generated Japanese
RNN	9.1	10.3	-	-
Transformer	6.8	8.2	3.5	5.1

BLEU and METEOR scores of text-to-text translation

Model	Syntactic similar				Syntactic distant			
	En to ES		Ja to Ko		En to Ja		Ja to En	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
RNN	45.9	62.1	46.3	63.8	42.0	58.4	43.4	59.8
Transformer	47.1	66.4	48.2	67.3	43.2	59.8	45.1	61.0

- ✓ The Transformer model outperformed the RNN model on ASR and text MT tasks.
- ✓ The syntactic similar language translation achieved a higher score than syntactic distant language.

Experiment Results

We evaluate Speech translation performance to compare.
- ✓ Proposed and Multi-task method.
- ✓ RNN and Transformer model.
- ✓ Syntactically similar and distant language.

BLEU and METEOR scores of speech-to-speech translation

Model	Syntactic similar				Syntactic distant			
	En to ES		Ja to Ko		En to Ja		Ja to En	
	BLEU		METEOR		BLEU		METEOR	
Baseline: Cascade (RNN)	38.9	47.7	38.7	49.1	32.5	44.2	32.0	43.2
Baseline: Cascade (Transformer)	41.3	52.1	41.0	51.1	34.1	45.2	35.0	45.3
Multi-task (RNN)	38.8	48.2	39.1	49.9	33.2	45.5	34.2	45.0
Multi-task (Transformer)	43.1	58.8	42.5	58.3	36.9	52.6	38.3	48.4
Transcoder(Transformer)	44.0	59.3	42.9	58.8	40.6	56.6	41.0	55.8

✓Our proposed approach outperforms Multi-task speech translation.

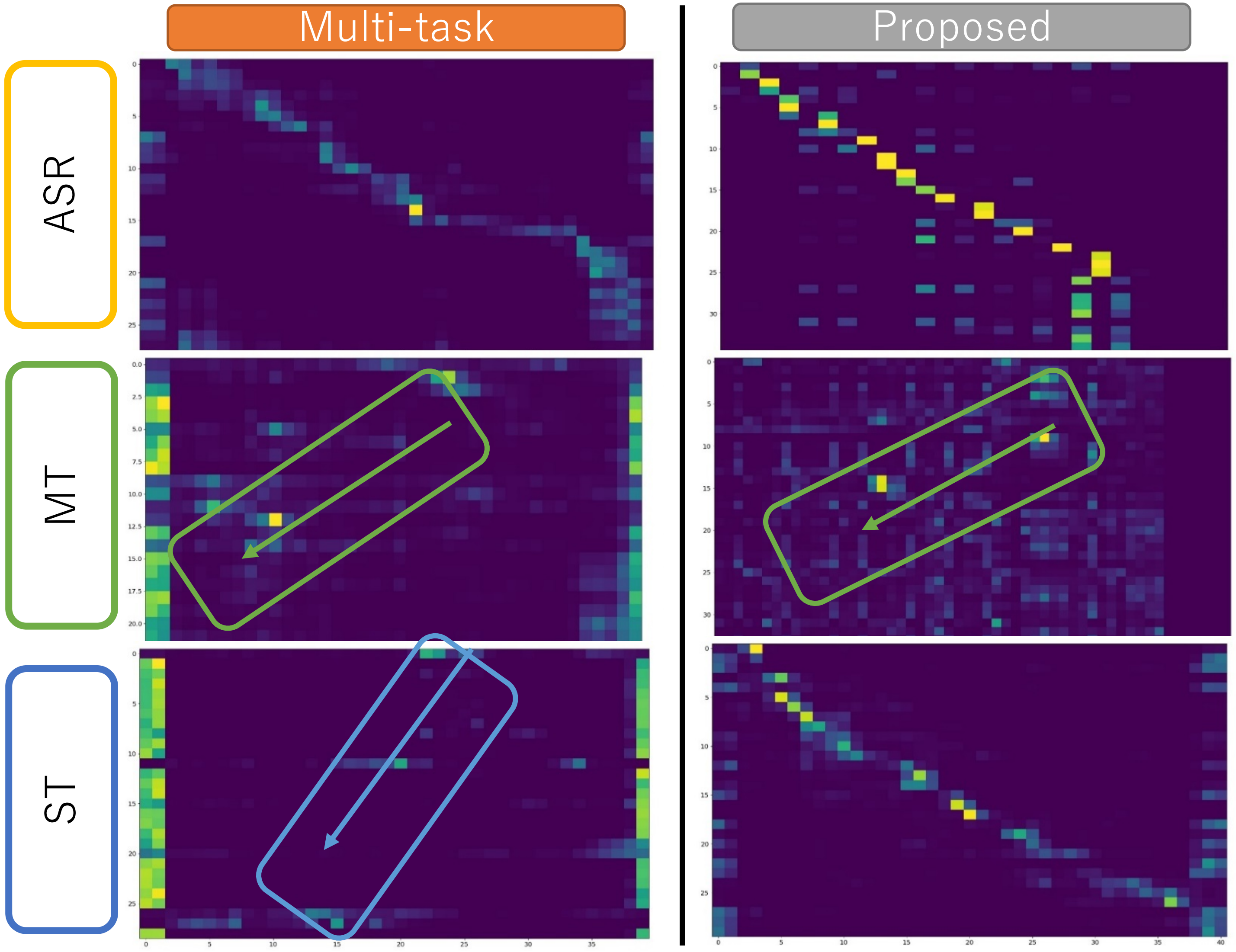
✓The proposed method with Transformer further improved performances.

Analysis of the proposed and Multi-task methods

- ✓ Both methods show monotonic attention alignments at ASR tasks. However, the Multi-task model's attention alignment is not clear.

- ✓ Both MT attention alignments has a similar direction, as shown in the arrow.

- ✓ The proposed method shows monotonic attention alignments again.
- ✓ The Multi-task method shows week attention alignments similar to MT but not precisely the same.



Conclusion

- In this research, we propose the Transformer-based speech-to-speech translation with a transcoder that can pass the context information for each process.
  - Our proposed model improves BLEU and METEOR scores compared with the Multi-task model on syntactically distant language.
- Our proposed model learns the end-to-end speech translation step by step.
  - ASR and MT decoders support alignment at each step.
- Our proposed model shows the best performance on both syntactically similar and distant language pairs.

Thank you for your listening.